

YOUR COURSE. YOUR EXAM PREPARATION.

Data Science Fundamentals

A detailed roadmap for Weeks 1-5

Understand the idea. Work the numbers.
Read the code. Defend the conclusion.

Based on 279 pages of lecture and lab PDFs, two notebooks, and the Week 3 lab discussion. Includes a 14-day plan, worked examples, complete Week 4 calculations and a 100-mark practice exam with answers.

New Uzbekistan University
Course materials: Mohammad Sabokrou
Prepared as an independent study companion

How to use this guide

Prepared for understanding, calculation, code interpretation and defensible written answers. This guide aims to help you prepare for the highest marks; it cannot predict the exact exam or guarantee a score.

What is covered: the five lecture decks, Lab 1, Lab 2, and the Week 4 lab instructions. The six PDFs contain 279 pages. The two notebooks contain 57 cells. Your earlier Week 3 lab discussion supplies additional practice context, clearly distinguished from the newly uploaded files.

What is not supplied: the course CSV datasets, your personalised lab ID/seed/output, the Week 3 notebook file, a separate Week 5 lab, the textbook chapters, later lecture decks, or an official exam paper. Course-wide example figures below are reported by the slides; personalised results must come from your own run. The fixed Week 4 lab calculations were independently recomputed.

How to use this guide: first read the roadmap and answer method; then study Weeks 1-5 in order. Attempt the practice exam before reading its answer key. Use the final checklist to decide what to revise.

Contents

1. What the course actually assesses	4
2. Week 1: asking the right question and counting the right thing	6
3. Week 2: describing data without hiding its structure	10
4. Week 3: cleaning decisions, missingness and audit	14
5. Week 4: representation, prediction and leakage	18
6. Lab 4: complete combined worked example	22
7. Week 5: EDA, confounding and Simpson's paradox	25
8. Calculation and code reference	29
9. Practice exam: 100 marks	31
10. Practice exam: answers and marking guide	33
11. Revision traps, writing templates and readiness	35
12. Source map and coverage checklist	37

1. What the course actually assesses

Source: W1 PDF pp. 17-20, 52, 56-60; L1 marking table; L2 introduction and submission checklist; L4 pp. 1, 7.

1.1 The exam strategy follows the assessment

The Week 1 slides describe the exams as written, in person, and applied: you are shown a table, result, short code excerpt or chart and asked to judge it. They explicitly say the papers do not ask for definitions alone. Learn terms so that you can use them to explain a decision; memorising a glossary is insufficient.

A strong answer does four things: makes the requested decision, cites the relevant numerical or contextual evidence, explains the mechanism, and names the remaining limitation. For example: "Keep the verified 47-year-old student. Their age exceeds an IQR fence, but source verification places them in the target population. Deleting them changes who is counted and can lower the reported withdrawal rate without improving any student's outcome."

The assessment weights reported in W1 are: midterm 20% in Week 6; final exam 25%; weekly labs 15% (best 9 of 12); Data Quality Report 10%; EDA and Visualization Brief 10%; Modelling Assignment 10%; capstone and dashboard 10%. These are the supplied slide schedule, not independently checked current administrative instructions. Follow any later announcement from your lecturer.

Scope warning: the uploaded pack is a strong foundation for Weeks 1-5 and preparation for the scheduled Week 6 midterm. It is not complete preparation for an exam covering Weeks 6-12. Confirm the assessed scope with the lecturer.

1.2 A reusable answer method

1. **Question and claim:** what is being asked? Description, comparison, prediction or causal effect?
2. **Population and row:** who should the conclusion concern, and what does one record represent?
3. **Denominator and units:** how many eligible records, how many valid measurements, and in what units?
4. **Method and evidence:** show the calculation or source evidence that supports your choice.
5. **Interpretation:** translate the result into the context rather than repeating a number.
6. **Limit or next check:** identify the specific uncertainty that could change the decision.

For short questions, compress this into two or three sentences. Do not write all six labels mechanically if the question asks only for a calculation. Always answer the requested part first.

Weak: "The data are biased."

Stronger: "Evening students are 20/100 of invitees but only 8/80 of respondents. Complete-case analysis therefore halves their representation from 20% to 10%. The pooled respondent mean may not represent the invited cohort."

1.3 A 14-day preparation plan

This is a suggested plan, not an official course requirement. Allocate about two focused hours per day; extend the calculation sessions if they are new to you.

Day	Focus	Evidence that you have learned it
1	Week 1: questions, claim ladder, populations and proxies	Repair three vague questions; identify what each claim needs
2	Week 1 + Lab 1: rows, keys, joins and denominators	Predict merge size and explain each TODO without notes

Day	Focus	Evidence that you have learned it
3	Week 2: mean, median, variance and SD	Complete the eight-number exercise by hand with units
4	Week 2: quartiles, boxplots, shape and z-scores	Explain fences versus whiskers and solve three z-score tasks
5	Lab 2: apply summaries to a decision	Write a 150-word recommendation using your own output
6	Week 3: quality, conversions and duplicate decisions	Build a decision table with evidence and alternatives
7	Week 3: missingness, sensitivity and audit	Explain MCAR/MAR/MNAR; calculate representation and bounds
8	Week 4: distances, scaling, log and encoding	Reproduce the A/B/NEW example and the future value 1.5
9	Week 4: availability, leakage and pipeline	Audit a broken workflow; transform the future vector by hand
10	Lab 4: combined student exercise	Reproduce saved statistics, distances and test error
11	Week 5: EDA, rates, plots and unknown outcomes	Choose appropriate plots and compare rates with denominators
12	Week 5: confounding and Simpson's paradox	Reconstruct 66% versus 74% and the common-mix rates
13	Closed-notes practice exam in this guide	Mark every answer and identify why marks were lost
14	Repair weak areas, then a mixed retest	Answer changed examples without copying remembered numbers

Daily pattern: 15 minutes recall yesterday from memory; 35 minutes study; 40 minutes solve or interpret; 20 minutes explain aloud or in writing; 10 minutes record mistakes. Revisit a difficult idea after one, three and seven days. These timings are recommendations, not a scientifically guaranteed schedule.

If you have only three days: day 1 covers Weeks 1-2 and their labs; day 2 covers Weeks 3-4; day 3 covers Week 5 and mixed timed practice. Prioritise row meaning, denominators, SD/quartiles, cleaning evidence, leakage and weighted rates. This compressed plan trades depth for coverage.

1.4 Study actively

Before running code: predict the number of rows, missing values, direction of change or nearest student. **After running:** cite what happened and explain any revision. Keep your original prediction visible in assessed notebooks.

Maintain an error log with four columns: question, my error, correct reasoning, new example. Categorise errors as arithmetic, denominator, unit, assumption, timing or overclaim. A repeated conceptual error deserves more attention than another reread of a familiar slide.

Use the labs as oral-exam practice: cover the output and explain what each line should do. Change the example: a repeated ID may be a copy in one table and a valid second week in another. Transfer is the skill the exam descriptions emphasise.

2. Week 1: asking the right question and counting the right thing

Sources: W1 PDF pp. 3, 13-17, 22-30, 32-53, 57-60; L1 Parts 1-5.

2.1 Data science, AI, machine learning and deep learning

Data science uses data to answer questions, quantify uncertainty and support decisions. AI is the broader aim of building systems with behaviours associated with intelligence. Machine learning is a way to learn patterns from data; deep learning is a family of machine-learning methods based on neural networks. Data science can use these methods, but cleaning a registry or comparing distributions can be valuable without a predictive model.

The Week 1 gallery illustrates transferable lessons: maps can expose spatial patterns; failure-only records omit a necessary denominator; a search-based prediction system needs monitoring as behaviour changes; operational or satellite records may answer a later question; fraud or screening thresholds trade different mistakes; recommendation, travel-time, language and irrigation systems all depend on what their recorded signals mean. Focus on these reasoning lessons rather than memorising historical dates.

2.2 Frame a question with four components

Population: who the result is about. **Variables:** what is actually measured. **Period:** when measurements and outcomes apply. **Purpose:** the decision the answer is intended to support. For a comparison, also specify the comparator.

"Students seem disengaged" is incomplete. A better question is: "Among the 2026 entry cohort, are recorded submissions during teaching weeks 1-4 lower than for the 2025 cohort at the same stage, and is the difference concentrated in particular programmes so support can be targeted?"

This operationalises engagement using submissions. It does not establish that submissions fully measure learning. The period, collection process and comparison must be comparable across cohorts.

2.3 The claim ladder

Claim	Example	Evidence required and limit
Descriptive	What withdrawal rate was recorded last year?	Correct count, eligible denominator, relevant population
Comparative	Do recorded pass rates differ by programme?	Comparable groups, counts, units and uncertainty; no automatic causal claim
Predictive	Can early signals predict later withdrawal?	Information available at prediction time and honest held-out evaluation
Causal	Would providing support reduce withdrawal?	A suitable design and identification assumptions, such as a defensible randomised comparison

A correlation is evidence of association. It can motivate a predictive model, but prediction performance still needs testing on unseen cases. It does not establish that forcing the input to change would change the outcome.

In W1, logins and mean grades are reported as correlated at about 0.68 for 1,200 students. Entry score is associated with both. Possible explanations include a causal effect, reverse direction and common causes such as preparation. Those correlations alone do not prove which explanation is correct or how much of the association each explains.

Confounder: in the causal question being considered, a pre-existing common cause can influence both the exposure and outcome. A variable merely correlated with both is a candidate to investigate, not automatic proof of a causal structure.

2.4 Tables, rows, columns and keys

A CSV is a text representation of a table. A DataFrame is a table that code can manipulate. A column describes an attribute; a row describes one observational unit. The **grain** or row definition says exactly what that unit is. A key uniquely identifies it; a composite key combines columns.

Course table	Slide-reported rows	One row and candidate key
students.csv	1,200	One student; student_id
enrolments.csv	4,665	One student-course registration; student_id + course_code if unique in the stated period
activity.csv	16,165	One student-week; student_id + week if the export has one record per pair
outcomes.csv	1,200	One student's early features and withdrawal outcome
feedback.csv	732	One text comment; comments need not equal distinct students
platform_daily.csv	364	One day of platform sessions
students_dirty.csv	1,205	A damaged student export; inspect extra records before treating them as people

If repeated course attempts or terms exist, a registration key may also require term or attempt. A key is justified by the process and verified against the data, not guessed from a column name.

`len(enrolments)` counts registrations. `enrolments['student_id'].nunique()` counts distinct recorded IDs. Neither alone proves that every expected student is present; compare with the roster. Multiple registrations give some students more weight in a row-level average.

W1 illustrates this with part-time representation: approximately 11.9% of registration rows versus 14.4% of students. The 2.5 percentage-point difference comes from unequal numbers of registrations per student. More rows do not fix an incorrect counting unit.

2.5 Measurement meaning determines valid arithmetic

Type	Examples	Meaningful operations
Nominal category	Programme, study mode	Counts, proportions, mode; no intrinsic order or mean
Ordinal category	Satisfaction 1-5; low/medium/high workload	Order, category distribution, median under an appropriate convention; equal gaps are not guaranteed
Discrete numerical	Number of logins, submissions	Numerical summaries when the counts have the stated meaning
Continuous numerical	Duration, measured score on a sufficiently fine scale	Numerical summaries with units and a suitable interpretation
Identifier	Student ID	Match, check uniqueness, count distinct; arithmetic averages have no useful interpretation

The slides treat scores as numerical measurements. A score recorded only as whole points is technically discrete; the important exam distinction is a quantitative measurement versus an arbitrary category code. Its stored Python type does not settle its meaning.

Keep identifiers as text to preserve leading zeros. A numeric-looking programme code does not become a meaningful measurement. Satisfaction means can be reported with an explicit equal-step convention, but show category frequencies too; an average can conceal polarisation.

2.6 Population, sampling frame and observed sample

The **target population** is everyone the conclusion should describe. The **sampling frame** is the list from which participants can be reached or sampled. The **observed sample** is who is actually recorded. For a platform survey these could be all enrolled students, students with active accounts, and those who reply.

Coverage error arises when the frame excludes parts of the target population. **Nonresponse** occurs when included or invited people do not provide an answer. A larger convenience sample does not repair systematic exclusion. Random sampling is a selection design, not a synonym for "a lot of data."

The Week 1 slides assign additional sampling strategies as textbook reading, but the textbook is not attached. Know the basic contrast: a probability sample uses a known selection mechanism; convenience selection is driven by access. More advanced sampling designs should be checked against your assigned reading if they are in scope.

2.7 Operational definitions, proxies and ethics

An operational definition is the explicit rule that turns a concept into recorded values. A proxy is a measurable stand-in for something not directly measured. Platform minutes may include an idle browser tab; low online activity may reflect offline study, work or caring responsibilities.

Reliability concerns consistency of measurement under comparable conditions. **Measurement validity** concerns whether it captures the intended concept. This is a different use of "validity" from a Week 3 range check. A consistently recorded login count can be reliable while being an incomplete measure of engagement.

Consider who benefits, who bears false alarms or missed cases, whether less intrusive data would suffice, and whether affected people can understand and challenge a decision. A sensitive attribute might be useful for auditing unequal outcomes without being an appropriate input for an individual decision. Predictive usefulness alone does not settle that choice.

2.8 Joins: predict the result before running it

A join combines records using keys. A **left join** retains left-table rows and adds matches; missing right-side matches become missing values. An **inner join** retains matching keys only and can silently narrow the population. Cardinality describes the number of possible matches: one-to-one, one-to-many, many-to-one or many-to-many.

Joining enrolments on the left to a unique student table is many-to-one. If the right key is unique, the left join should retain 4,665 registration rows. Duplicated student keys can multiply rows. A left row matching two right rows appears twice; with repeated keys on both sides, matches can multiply within each key.

```
joined = enrolments.merge(
    students[['student_id', 'programme', 'study_mode']],
    on='student_id', how='left', validate='many_to_one',
    indicator=True
)
assert len(joined) == len(enrolments)
print(joined['_merge'].value_counts())
```

`validate='many_to_one'` checks uniqueness on the right; it does not prove all IDs matched or that the source values are accurate. The merge indicator distinguishes unmatched records from a matched record whose programme itself is missing. For the provided Lab 1, a programme-not-missing assertion

is also used as a practical check under its data assumptions.

2.9 Lab 1: what you should be able to explain

The four TODO replacements are `students`, `one_student_table`, `'study_mode'`, and `'programme'`. Understand why each belongs there.

```
one_student_table = students
joined = enrolments.merge(
    one_student_table[['student_id', 'programme', 'study_mode']],
    on='student_id', how='left', validate='many_to_one'
)
pass_rate_by_mode = (
    joined.groupby('study_mode')['passed']
    .agg(n='size', pass_rate='mean')
    .assign(pass_rate=lambda x: 100 * x['pass_rate'])
)
```

For a complete Boolean/0-1 outcome, its mean equals passes divided by registrations. `size` counts group rows; `count` counts nonmissing values. If `passed` has missing values, the mean excludes them while `size` includes them: explicitly report the two denominators. The Lab 1 code assumes the provided pass outcomes are complete.

The final reflection must explain why the programme with the highest pass rate is not automatically the best teaching programme: preparation, course selection, grading thresholds and missingness can differ. Your chart is a descriptive comparison of registrations, not an intervention evaluation.

2.10 Reproducibility and provenance

The analysis lifecycle is iterative: frame, acquire, understand and clean, explore, model if needed, evaluate, communicate, and monitor. A discovery can send you back to an earlier stage.

A data dictionary records names, meaning, type, units, allowed values and missingness rules.

Provenance records source, collection context, dates, permissions where relevant, and transformations.

A reproducible notebook combines purpose, code, output and interpretation. Restart and run all; hidden state and execution order can otherwise make a correct-looking notebook impossible to reproduce.

3. Week 2: describing data without hiding its structure

Sources: W2 PDF pp. 2-60; L2 Sections 1-8. The personalised Lab 2 has 240 synthetic students, unlike the 1,200-student lecture dataset. Do not mix their numbers.

3.1 The summary contract

Before calculating, state the observational unit, measurement unit, target population, valid n , missing count and distribution shape. Then choose a centre and spread suited to the question. A summary compresses data: it necessarily omits individuals, tails, modes, groups and collection context. Say which omission matters to the decision.

`len(x)` counts entries including missing ones; `x.count()` counts observed values; `x.mean()` normally ignores missing values. A mean from 916 respondents is not automatically the mean for all 1,200 invited students. The W2 lecture example has 284 missing responses, about 23.7%.

3.2 Mean, median and mode

The **mean** is sum divided by valid count. It uses every value and is the balance point: deviations from it sum to zero. The **median** is the middle sorted value, or the average of the middle two for an even-sized numeric list. It depends on position and is less sensitive to extreme magnitudes. The **mode** is the most frequent value or category; there may be several modes.

Mean and SD are often useful for a roughly symmetric, single-group numerical distribution. Median and IQR often describe a typical observation in a long-tailed distribution. For a category, report counts or proportions and possibly the mode. A total is related to the mean, not generally to the median.

W2 reports platform-use mean 1,592 and median 1,426 minutes per student over the term. The mean is pulled upward by high users. For a typical student, median plus IQR is natural. For total recorded workload, sum or valid n times the observed mean is appropriate. Multiplying an observed mean by a larger population count with missing values introduces an assumption.

3.3 Variance and standard deviation

Let n be the number of observed values, and $\bar{x}$ their mean. A deviation is x_i minus $\bar{x}$. Squaring makes negative and positive deviations contribute positively.

Sample variance: sum of squared deviations divided by $n-1$. **Sample SD:** square root of sample variance. **Population-style variance:** divide by n when describing the entire listed population or following a specified scaling convention.

The $n-1$ correction accounts for estimating the mean from the same sample. Under the usual independent random-sample assumptions it makes sample variance unbiased for population variance. It does not guarantee every sample estimate is too small before correction, and sample SD itself is not exactly unbiased.

Variance has squared units, such as points squared. SD returns to points. SD describes dispersion of observations; it is not the standard error of a mean and not a typical prediction error.

Pandas `Series.std()` defaults to `ddof=1`; NumPy's ordinary array SD defaults to `ddof=0`. With missing values, also choose matching missing-value handling before comparing implementations. One observed value is insufficient for sample SD; a constant group has SD zero.

3.4 Worked eight-number calculation

Additional practice example, not your personalised Lab 2 output: **35, 40, 45, 50, 60, 65, 70, 75** points.

Quantity	Calculation and answer
Sum and n	440 points; $n = 8$
Mean	$440 / 8 = 55$ points
Median	$(50 + 60) / 2 = 55$ points
Deviations	-20, -15, -10, -5, 5, 10, 15, 20
Squared deviations	400, 225, 100, 25, 25, 100, 225, 400
Sum of squared deviations	1,500 points squared
Sample variance and SD	$1,500 / 7 = 214.286$ points squared; SD = 14.639 points
Population variance and SD	$1,500 / 8 = 187.5$ points squared; SD = 13.693 points

Show the intermediate steps in a hand-calculation question. Round the final answer, not each intermediate deviation or fitted parameter.

3.5 Quartiles, percentiles and conventions

Sort first. Q1 is the 25th percentile, Q2 the median, and Q3 the 75th percentile. The five-number summary is minimum, Q1, median, Q3 and maximum. $IQR = Q3 - Q1$ and describes the width of the middle half.

In Lab 2's eight-number exercise, use **median of halves**: find the median of the lower four and the upper four. For our practice list: $Q1 = (40+45)/2 = 42.5$; $Q3 = (65+70)/2 = 67.5$; $IQR = 25$ points. Five-number summary: 35, 42.5, 55, 67.5, 75.

For the larger Lab 2 dataset, pandas uses **linear interpolation**. On the same eight values, it gives $Q1 = 43.75$ and $Q3 = 66.25$. These differ because the conventions locate quartiles differently in finite samples, not because one analyst necessarily made an arithmetic error. State the convention.

A percentile describes rank, not ability or causation. With ties, the exact empirical rank convention matters; the lab explicitly uses the fraction at or below the target value.

3.6 Range, fences and boxplots

Range = maximum minus minimum; distinguish this width from the endpoint interval. It depends only on two values. When you add observations without removing any, the range cannot shrink, although it can stay unchanged. SD and IQR can rise or fall as new data arrive; they do not have the same monotonic property.

Lower fence = $Q1 - 1.5 \text{ times } IQR$. **Upper fence** = $Q3 + 1.5 \text{ times } IQR$. A typical Tukey boxplot has a box from Q1 to Q3, a median line, and whiskers at the most extreme observed values still inside the fences. The whiskers are not the numerical fence locations. Points beyond the whiskers are candidates for review.

Using lecture quartiles 58.3 and 73.6, $IQR = 15.3$, and fences are 35.35 and 96.55 points. A score of 99 is flagged but still lies within a documented 0-100 scale. Verify the record before deciding what to do.

In the eight-number practice list, multiplying 75 by 10 changes the mean from 55 to 139.375 and the range from 40 to 715; sample SD becomes about 247.031. Median remains 55 and median-of-halves IQR remains 25 because the middle positions used have not changed. This illustrates robustness in this example, not immunity to every possible contamination pattern.

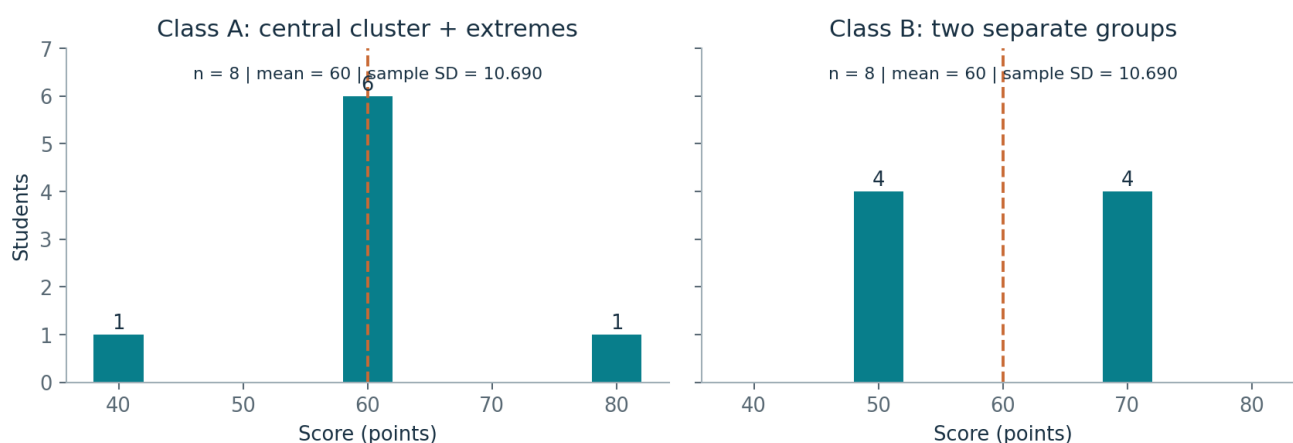
3.7 Shape, histograms and hidden groups

Inspect centre, spread, tail direction, peaks or gaps, and unusual values. A right tail extends toward large values; a left tail extends toward small values. Mean greater than median often accompanies a right tail, but it is a clue rather than a mathematical proof of shape. Similar mean and median or near-zero skewness does not establish normality.

A histogram groups numbers into intervals. Few wide bins may conceal two peaks; many narrow bins may exaggerate noise. Changing bins changes the picture, not the underlying observations or their mean. Compare reasonable choices and label units and valid n. For satisfaction 1-5, one bar per category preserves its discrete ordinal structure.

A bimodal distribution can put the overall mean between two groups where few people lie. Disaggregate using a meaningful variable and report each group's n, centre and spread. Do not infer that group membership caused the difference.

The exact Lab 2 constructions use offsets [-20,0,0,0,0,0,20] and [-10,-10,-10,-10,10,10,10,10]. Both have mean offset 0 and sample SD $\sqrt{800/7}$, about 10.690. The first has six central students and two extremes; the second has two groups. They can require different tutoring plans despite identical mean and SD.



Additional visualisation of the exact Lab 2 offset construction. Both classes have the same mean and sample SD, but their students occupy different parts of the scale.

3.8 The normal-model rule

Under a normal model, approximately 68%, 95% and 99.7% lie within one, two and three SD of the mean. These are model probabilities, not guarantees for a finite sample. A symmetric distribution can be bimodal and unsuitable for this approximation. One observed proportion near 95% does not prove normality.

For mean 65.9 and SD 10.9, the intervals are 55.0-76.8 for one SD, 44.1-87.7 for two, and 33.2-98.6 for three. Say "approximately" and state the roughly normal assumption. Do not automatically apply this interpretation to strongly right-skewed minutes.

3.9 Z-scores: relative to which group?

$z = (\text{value} - \text{reference mean}) / \text{reference SD}$. It is unitless and says how many reference SDs the value is above or below the reference mean. Positive does not automatically mean good. Computing z does not require a normal distribution; turning z into a normal-model percentile does.

For score 99, cohort mean 65.86 and SD 10.94, z is about 3.03. The same score can have different z-values against another programme because the reference mean and SD differ. State the population and the SD convention.

If all values are identical, the reference SD is zero and the usual z formula is undefined. Standardising arbitrary programme codes or IDs gives numbers without a meaningful quantitative interpretation. A

standardised column has mean approximately zero and SD one when evaluated on the same reference data with the same convention, apart from numerical precision.

Distance versus rank: in the lab's minutes $[1, 2, \dots, 9, M]$, M ranges from 46 to 66. The mean is $(45+M)/10$, greater than 9, so 9 is below the mean. Nevertheless 9 of the 10 observations are at or below 9: it is at the 90th percentile by the stated empirical convention. The large tail value separates mean position from rank.

Specification versus statistical unusualness: bottle reference mean 100 mL and SD 2 mL, with an agreed acceptance interval 94-106 mL. A 104 mL bottle has $z=2$ and is accepted; a 108 mL bottle has $z=4$ and is rejected. Acceptance comes from the specification, not from an automatic z cutoff. Preserve the rejected bottle's measurement for process analysis.

3.10 Lab 2 completion map

Section	What you must demonstrate
1	Observational unit; response-rate versus observed-mean denominator; ordinal/category/ID meaning
2	Eight-number arithmetic, squared units, n versus $n-1$ and software defaults
3	Stated quartile convention; sensitivity to a changed maximum; impossible score versus valid high usage
4	Four plot descriptions; binning; why zero skewness does not prove normality
5	Fences versus observed whiskers; source checks; group summaries; same mean and SD but different needs
6	Observed proportions versus the normal-model rule, without treating approximation as proof
7	Reference-group z -scores; zero SD; percentile contrast; bottle specification
8	Typical-use versus total-workload recommendation with own case code, numbers and limitations

The notebook uses your ID to generate its seed and case code. Keep the same ID and use your own results. The lecture's 1,200-student figures do not replace your 240-student lab results. Restart and run all before submission, and do not delete a high-use student merely because a fence flags them.

4. Week 3: cleaning decisions, missingness and audit

Source: W3 PDF pp. 4-55. The 14-row registry example below comes from the Week 3 lab code you previously pasted in this conversation; it is separate from the lecture's smaller examples.

4.1 Quality means fitness for the task

A blank phone number prevents an urgent phone call. It does not prevent counting a known programme or analysing an available score. Judge quality relative to the intended task, rather than treating every incomplete row as useless.

Dimension	A concrete question or check
Completeness	Are expected records and required cells present? Compare with a roster as well as counting blanks.
Validity	Does each value satisfy the documented type, category, range or format?
Consistency	Do related fields and representations agree, such as start before end and approved programme aliases?
Uniqueness	Is the intended key unique at the stated row grain?
Accuracy	Does the stored value agree with the correct measurement for that person and event?
Timeliness	Is the information current enough for this purpose, considering event time and snapshot time?

A score of 67 and a score of 76 both pass a 0-100 range check. If the marked exam records 76, only the second is accurate. Assertions cannot establish a source fact they never check. A timestamp says when something was updated, not that a phone number still works.

4.2 Inventory before editing

Preserve the raw export. Record rows, distinct keys, expected IDs absent, missing cells, stored types, allowed values and ranges. Note which calculations use records versus valid measurements. A table may have more rows than students and still omit an expected student; three counts must be checked separately.

A defensible cleaning plan has the issue, affected record or value, proposed action, source evidence, and risk of an alternative. An audit log distinguishes rows removed, cells standardised, values flagged and records retained. These units must not be added together as if all were row removals.

4.3 Meaning comes before conversion

Trim outside spaces and map approved aliases: 'biology', 'Biology' and 'BIO' may become Biology when the dictionary confirms the equivalence. Bioengineering stays separate. Similar spelling alone does not authorise merging concepts.

Percentages require a declared target representation. If a proportion column uses 0-1, '67%' becomes 0.67. If the entry-score dictionary uses 0-100 points, a verified '67%' representation becomes 67 points. W3's general conversion slide and its score exercise use different target units; neither rule should be applied blindly to every column.

'00124' is an ID and should keep its leading zeros. '1,200' needs a locale/separator convention. '03/04/2026' is 3 April under day/month/year and 4 March under month/day/year. Both parse successfully; source evidence chooses the meaning. In your earlier registry lab, the explicit format is

%d/%m/%Y.

`pd.to_numeric(..., errors='coerce')` makes failed conversions missing. It does not infer zero or recover a true exam mark. Keep the raw text and distinguish originally blank from unparsed. 'seventy' contains information but remains unresolved in this lab until source verification supplies an approved numeric value.

4.4 Repeated records need different actions

Confirmed copy: remove the extra copy for a current registry. **Verified correction:** use the authoritative revision and retain the earlier record in the audit history. **Unresolved conflicting values:** flag and investigate rather than arbitrarily taking the first or latest. **Repeated observations:** retain different weeks, courses or events and use the appropriate composite key.

'Keep the latest' is justified only when the timestamp and source process identify a genuine superseding correction. A later upload can be stale information, a different event or a less authoritative source.

Worked W3 question, PDF p. 21: after removing the confirmed copy, Biology contains a score of 70 and an unresolved score of either 60 or 80. Candidate Biology means are 65 and 75. Reporting 75 as settled is unjustified until the conflicting version is verified. Bioengineering's 90 belongs to a different programme.

4.5 Zero, missing and not applicable

Zero lab-use hours means an enrolled student's observed use was zero. Missing hours means use is unknown. Not enrolled means the lab-use question may not apply; those students are outside a target population defined as enrolled lab students. Keep enrolled students with missing hours in the population count and report how many measurements are unavailable.

A missing marker may be an empty string, NaN, NA, a documented sentinel or an explicit not-applicable code. They are not automatically interchangeable. For example, the earlier lab's `eq('')` detects exact blank strings, not every possible kind of missing data in a general export.

4.6 MCAR, MAR and MNAR are mechanisms

Mechanism	What must be true	Course story
MCAR	Missingness is independent of observed and unobserved values relevant to the mechanism.	Answers are independently lost with equal probability regardless of class or rating.
MAR given class	Once known class is accounted for, reply probability does not depend on the missing rating.	Day satisfied/dissatisfied both reply with probability 90%; Evening both with probability 40%.
MNAR	Missingness still depends on the missing value after conditioning on the observed information used.	Within Evening, satisfied reply with probability 40%, dissatisfied with probability 10%.

MAR does not mean all students are equally likely to reply, and it does not mean complete-case estimates are automatically unbiased for the full population. MCAR removes systematic selection under the stated process but still leaves finite-sample variability and fewer observations.

Observed response rates do not identify the mechanism. Different rates can occur by chance or through dependence on known or unknown factors. Equal rates can hide strong bias: in two groups with half rating 1 and half rating 5, only the 5s might answer in one group and only the 1s in the other. Both have 50% response but observed means 5 and 1, despite true means both 3.

Follow up a random subset of nonrespondents using another channel. Their responses may inform the missingness story, but the people still not responding can differ, and contact mode can affect answers. State the remaining uncertainty.

4.7 Calculate representation explicitly

In the W3 lecture example, 80 Day students yield 72 replies and 20 Evening students yield 8 replies. Response rates are $72/80=90\%$ and $8/20=40\%$. Missing fractions are $8/80=10\%$ and $12/20=60\%$.

Evening students make up $20/100=20\%$ of invitees but $8/80=10\%$ of respondents. Dropping missing ratings changes group composition. Keep clear the distinction between a within-group response rate and a group's share of all replies.

With respondent means 4 for Day and 2 for Evening, the pooled observed mean is $(72*4+8*2)/80 = 3.8$. If responders represent nonresponders within each class, a class-size weighted estimate for all students is $(80*4+20*2)/100 = 3.6$. That estimate depends on an assumption; it is not a newly observed mean.

If the 12 missing Evening ratings average 1 while all Day ratings average 4, the all-student mean becomes $(80*4+8*2+12*1)/100 = 3.48$, below a 3.5 target. This is why the respondent mean alone cannot establish the target claim.

4.8 Four responses to missing data

Recover: copy a verified score from the exact student's exact exam, recording provenance. **Keep missing:** use other valid fields and report coverage. **Exclude for a stated analysis:** define the analysis population, number excluded and potential selection bias. **Impute:** use an explicit estimate or model, mark replacements and report assumptions and uncertainty.

The Week 3 final registry stays unfilled. Week 4 later introduces imputation for predictive inputs fitted on training data. These serve different tasks. Neither treats an imputed value as observed truth.

In $[2,4,\text{missing},\text{missing},\text{missing}]$, the observed mean is 3 and sample SD is $\sqrt{2}$, about 1.414. Fill each blank with 3: the mean stays 3, numeric count increases from 2 to 5, but sample SD falls to $\sqrt{0.5}$, about 0.707. The number of observed answers remains two. Mean filling reduces visible spread and can distort relationships; it does not automatically represent uncertainty about the missing values.

4.9 Sensitivity analysis and bounds

Let N be all eligible people, n the observed count, T the observed sum, and m an assumed mean among the $N-n$ missing people. A hypothetical full-population mean is $[T+(N-n)*m]/N$. Evaluate several plausible m values without modifying the final observed dataset.

W3 pp. 33-34: $N=20$, $n=12$, observed mean=4, so $T=48$ and eight ratings are missing. Assuming $m=1$ gives 2.8; $m=3$ gives 3.6; $m=4$ gives 4.0; $m=5$ gives 4.4. On a 1-5 scale, full-population bounds are 2.8 to 4.4. A 3.5 target could be met or missed. These are logical bounds under the scale and roster assumptions, not confidence intervals.

Your earlier 30-student survey uses a different seed-dependent n . Recalculate using its actual output; do not copy 2.8-4.4 or the 100-student class example into that notebook.

4.10 Outliers, invalid values and the population

An IQR flag means unusual relative to the observed distribution. The dictionary determines validity; source evidence determines accuracy. Age 220 is unresolved invalid data. A verified 47-year-old student is an unusual valid member of the cohort. Age 21 could still be a typo for 31.

Flag invalid age as missing while preserving the student's usable programme and score. Correct a value only with evidence. Retain valid influential values, choose robust summaries when appropriate and report sensitivity as a comparison, not as a replacement for the intended-population result.

In W3 p. 43, all 100 students include 10 withdrawals. Five verified older students include four withdrawals. Removing them leaves $6/95=6.32\%$, versus $10/100=10\%$. Nobody's outcome improved; numerator, denominator and represented population changed. Lead with 10% if the target is all enrolled students, and show subgroup rates with their small sample sizes.

4.11 Your earlier registry: the fixed outcomes

Decision	Consequence
Start	14 rows, 12 distinct student IDs, zero absent expected IDs
Remove confirmed R13 copy of R01	13 rows
Use verified R14 correction instead of R02	12 rows
Flag R03's age 220	Still 12 rows; usable score retained
Parse scores while preserving original text	R06 originally blank; R10 unparsed 'seventy'; 10 valid numeric scores
Keep verified older student and Bioengineering	Population and programme meaning preserved
Final result	12 rows, 10 observed scores, 2 unavailable scores

If `drop_student` is selected instead, R03 is lost and the final count becomes 11; its valid score `scores[2]` is discarded. Restore `flag_value` and rerun the cleaning cell before final checks. The exact observed-score mean and version code depend on your own seed and are not inferred from the snippets.

Checks should cover key uniqueness, full expected ID set, ID text format, approved categories, observed value ranges, missing-status consistency, unchanged raw export and reconciled row removals. Passing them shows compliance with those rules; it does not prove every valid-looking score equals the actual exam mark.

5. Week 4: representation, prediction and leakage

Sources: W4 PDF pp. 2-42; L4 Tasks 1-7. All fixed worked values here were recalculated from the supplied examples. These toy rules are teaching devices, not deployable student-support recommendations.

5.1 Define the prediction problem

One row is one student. At the end of Week 4, inputs are information genuinely available at that time; the target is later withdrawal. A feature is an input measurement or derived summary. A label is the known target for a historical example. A classifier predicts a category such as Withdraw/Stay; predicting a numerical quantity is a different task.

ID is retained for tracking and joins, not included as an arbitrary numerical distance feature. Set membership (train/test/new) and later outcome must not be predictors. Historical outcomes may be training labels and test outcomes may score fixed predictions, but neither role permits leaking an unavailable outcome into inputs.

5.2 Nearest-student reasoning by hand

Past A: score 70, minutes 60, Stay. Past B: score 80, minutes 180, Withdraw. NEW: score 79, minutes 100.

Using score alone, the absolute gaps are 9 and 1, so the nearest student is B and the copied prediction is Withdraw.

Using Euclidean distance on both raw numbers, distance to A = $\sqrt{9^2 + 40^2} = 41$. Distance to B = $\sqrt{1^2 + 80^2}$, about 80.006. A is now closest, producing Stay. The minutes differences dominate the squared distance. Changing minutes to seconds would multiply those differences by 60 and their squared contribution by 3,600.

Scaling makes an explicit choice about feature representation and weighting. It does not reveal a uniquely correct distance or guarantee that prediction improves.

5.3 Standardisation, fit and transform

Fit means learn and save parameters from training data. **Transform** means apply the saved rule to an input without relearning. Split before fitting anything data-dependent.

For A and B only, score mean=75, population SD=5; minutes mean=120, population SD=60. W4 divides by n, matching the teaching scaler convention. This differs from W2's sample SD exercise and is intentional.

A transforms to [-1,-1], B to [1,1], and NEW to [0.8,-1/3]. Distance to A is $\sqrt{1.8^2 + (2/3)^2}$, about 1.919. Distance to B is $\sqrt{(-0.2)^2 + (-4/3)^2}$, about 1.348. B is nearest again. NEW must not influence the training mean or spread.

Use the saved training rule on training, test and future rows. Fitting a separate test scaler would put the two sets into different reference systems. Fitting one scaler on their union would leak test information into fitting.

5.4 Min-max scaling

Formula: $(x - \text{training minimum}) / (\text{training maximum} - \text{training minimum})$. With training values [4,12,16], save minimum=4 and maximum=16. The outputs are 0, $8/12=0.667$ and 1.

A future value 22 maps to $18/12=1.5$. That is correct under the saved reference. Training observations lie within [0,1]; future observations need not. Clipping to [0,1] is an additional information-changing policy, and refitting with the new value changes the reference. Decide and document policies before

evaluation.

If training maximum equals training minimum, the hand formula divides by zero. Treat a constant feature explicitly rather than pretending a meaningful scale can be estimated from no variation.

5.5 Log transformation and trees

For nonnegative counts, $\log_{1p}(x) = \ln(1+x)$: 0 becomes 0, 9 becomes about 2.303 and 99 about 4.605. The +1 permits zero. The transformation preserves order, compresses large gaps, and changes differences, ratios and units. It does not automatically make a variable normal or repair invalid records.

An ordinary threshold decision tree uses order rather than Euclidean distance. The rule $\text{score} < 75$ becomes $z < 0$ under the saved score scale, with identical decisions: 70 Stay, 80 Withdraw, 79 Withdraw. Thus scaling is often unnecessary for ordinary numerical tree splits. This is not a claim that every model or implementation behaves identically under every transformation.

Choose transformation based on the question, feature meaning and model. Standardisation uses training spread; min-max uses training endpoints; log compresses a long numerical tail. Preserve original measurements and the reason for transformation.

5.6 Encoding categories and unknown values

Programme is nominal. Assigning CS=0, Economics=1, Psychology=2 invents order and numerical distances. One-hot encoding with fixed column order [CS,Economics,Psychology] gives [1,0,0], [0,1,0] and [0,0,1]. Low<Medium<High is ordinal, but an ordering still does not guarantee equal gaps.

The lab chooses an ignore-unknown rule: Physics becomes [0,0,0]. That means no known category match, not missing all evidence about the person. Another defensible design could predeclare an Unknown category. Never silently change the fitted vocabulary or column order when a new student arrives. Monitor unknown categories and document the chosen rule.

One-hot representation avoids an invented rank, but its contribution to distance is still a modelling choice. Two different known categories differ in two indicator positions. An all-zero unknown differs from each known category in one. Interpret this policy when assessing a distance-based rule.

5.7 Engineer features with a source, unit and window

The sum of W1-W4 logins [3,4,2,5] is 14 logins, usable only after all four weeks are complete and accessible. It is not hours, attendance or direct proof of engagement.

Minutes per login: R has $200/10=20$; S has $100/5=20$. The same ratio conceals different totals. Keep totals if they matter. With zero logins, the ratio is undefined; use missing/None and a zero-login indicator. Investigate a combination such as positive minutes and zero logins according to source definitions instead of assuming a correction.

Elapsed time uses the correct event: at the 5 September 2026 snapshot, first registration on 1 September gives 4 days and on 4 September gives 1 day. A record-update timestamp is not necessarily first registration. Use documented formats and time zones where they affect the calculation.

5.8 Event time and availability time

Candidate feature for end-of-W4 prediction	Decision and reason
W1-W4 logins already visible	Use: completed and available by the cutoff
Programme historically recorded at enrolment	Use the historical version known then
Future exam date announced in W1	Use if the schedule was known then; a future event can have an available planned date
W3 quiz score first released in W5	Do not use: past event, unavailable result
Current GPA updated in W6	Do not use that later version as if it existed in W4
Whole-semester attendance	Do not use: includes activity after the cutoff

An event can happen before the cutoff while its measurement arrives after it. Conversely, a schedule for a future event can already be known. The prediction system's actual access time determines availability.

5.9 Three leakage routes

Future leakage: using information unavailable at the decision time, such as W5 logins in a W4 prediction. **Target leakage:** putting the answer or an answer-revealing field into the predictors. **Test leakage:** letting held-out examples influence preprocessing, feature choice or model selection.

Example: training values [2,4] have mean 3. Adding held-out [100] gives 35.333. Only 3 may be saved as the training mean. The problem exists even without reading test labels because test features have influenced fitting.

Do not put the same student's registrations on both sides when evaluating performance on new students. Such a split can share person-specific information. For a future-time prediction task, design the split to respect time too. The split must mimic the kind of generalisation claimed.

5.10 Pipelines preserve the fitted procedure

Safe order: define target and cutoff; separate training/test; fit preprocessing on training; transform each set with the saved rules; fit the prediction rule on training inputs and labels; evaluate the fixed procedure on held-out data. Use validation for choosing settings. Keep the final test for the final assessment rather than repeatedly repairing the model after seeing its score.

Numerical columns may take imputation then scaling. Categorical columns take encoding. Dates may produce elapsed-time features. IDs serve matching and tracking. A ColumnTransformer is a software mechanism for sending different column types through different routes; a Pipeline stores an ordered procedure.

If a training numeric column has missing values, learn its replacement from observed training values, fill training inputs, then learn the scaler on those completed training columns. Repeat the fitted steps without relearning on test/future rows. In cross-validation, the same principle applies inside each training fold; this is a useful extension of the boundary rule, not a new licence to use the validation fold for fitting.

5.11 A five-position vector by hand

Training scores [60,60,80,80] have mean 70 and population SD 10. Training logins [2,2,6,6] have median 4, mean 4 and population SD 2. Keep category order [CS,Economics,Psychology].

For a future student with score 78, missing logins and programmeCS: impute logins 4; $\text{score_z} = (78 - 70) / 10 = 0.8$; $\text{logins_z} = (4 - 4) / 2 = 0$; $\text{encodeCS} = [1, 0, 0]$. Final vector: **[0.8, 0, 1, 0, 0]**, meaning score_z, logins_z, CS, Economics, Psychology. No parameter is fitted from this student. This example prepares inputs; without a fitted prediction rule and labels, the vector alone is not a withdrawal prediction.

5.12 Error, accuracy and model complexity

Classification error = wrong predicted labels / evaluated cases. Accuracy = 1 - error when each evaluated case has one true and one predicted label and all are counted the same way. In the three-row lecture example, one mistake gives error $1/3 = 33.3\%$ and accuracy $2/3 = 66.7\%$.

Training error measures fit to familiar data. Test error measures performance on held-out cases under that sampling and split design. Perfect training performance is insufficient evidence about new students.

An underfit rule is too simple to capture important structure; an overfit rule captures details specific to training data. Increasing flexibility can reduce training error while worsening unseen-case error. This is a common pattern, not a mathematical requirement that every test error must exceed training error in every finite sample. Validation guides model choices; tiny test samples make performance estimates unstable.

6. Lab 4: complete combined worked example

Source: L4 pp. 6-7, Task 7. Calculations below use the fixed supplied split and all six requested feature positions. Keep full precision in code; displayed values are rounded.

6.1 Records, split and early activity

ID/set	Score; minutes	W1-W4 total logins; programme; outcome
S1/train	60; 60	2; CS; Withdraw
S2/train	65; 90	4; Economics; Withdraw
S3/train	80; 180	12; Psychology; Stay
S4/train	85; missing	16; CS; Stay
T1/test	62; 70	2; CS; Withdraw
T2/test	82; 170	14; Economics; Stay
N1/new	78; missing	10; Physics; unknown

All these early features are stated to be available at the cutoff. The four S rows fit the rules. T outcomes are used only after predictions; N1's true outcome remains unknown.

6.2 Saved parameters and imputation

Observed training minutes are [60,90,180], median 90. Fill S4 and N1 with 90 using that saved median. Pandas converts None in a numerical column to NaN: use `fillna`, not an `is None` check on the resulting cell.

Feature	Training mean after filling	Population SD (ddof=0)
score	72.5	10.307764064
minutes	105	45
early_logins	8.5	5.722761571

Minutes per login are S1=30, S2=22.5, S3=15, S4=5.625, T1=35, T2 about 12.143 and N1=9. S4 and N1 ratios use imputed minutes, so mark that provenance. All login totals are positive here, giving zero-login indicators of 0. These ratios are explanatory outputs; they are not additional coordinates in the specified prediction vector.

6.3 Six-position vectors

The feature order is [score_z, minutes_z, early_logins_z, CS, Economics, Psychology].

ID	Three scaled numbers	Programme indicators
S1	[-1.212678, -1.000000, -1.135815]	[1,0,0]
S2	[-0.727607, -0.333333, -0.786334]	[0,1,0]
S3	[0.727607, 1.666667, 0.611593]	[0,0,1]
S4	[1.212678, -0.333333, 1.310556]	[1,0,0]

ID	Three scaled numbers	Programme indicators
T1	[-1.018650, -0.777778, -1.135815]	[1,0,0]
T2	[0.921635, 1.444444, 0.961074]	[0,1,0]
N1	[0.533578, -0.333333, 0.262111]	[0,0,0]

6.4 Distances and predictions

Candidate	Distance to S1 / S2 / S3 / S4	Nearest and prediction
T1	0.295008 / 1.550604 / 3.752087 / 3.340821	S1; Withdraw
T2	4.114310 / 2.988968 / 1.486327 / 2.316751	S3; Stay
N1	2.539302 / 1.920892 / 2.271516 / 1.600129	S4; Stay

T1 and T2 are both correct: observed test error $0/2=0\%$, accuracy $2/2=100\%$. Two held-out cases provide very weak evidence about reliable performance. N1's prediction is reported separately; its unknown true outcome cannot be scored. Do not tune the rule after inspecting these test outcomes.

The final vector includes one-hot coordinates, not just the three scaled numerical values. A tie is broken by smaller training ID as specified. ID itself is never a distance coordinate.

6.5 Readable implementation

Run this after creating the `df` with the exact rows and columns supplied in Task 7. It follows the printed lab policy rather than choosing a new model.

```
import numpy as np

df['early_logins'] = df[['w1', 'w2', 'w3', 'w4']].sum(axis=1)
parts = {
    name: df.loc[df['set'].eq(name)].copy()
    for name in ['train', 'test', 'new']
}
train = parts['train']
saved_median = train['minutes'].median()
for part in parts.values():
    part['minutes_imputed'] = part['minutes'].isna()
    part['minutes'] = part['minutes'].fillna(saved_median)
    denom = part['early_logins'].replace(0, np.nan)
    part['minutes_per_login'] = part['minutes'] / denom
    part['zero_logins'] = part['early_logins'].eq(0).astype(int)

numeric = ['score', 'minutes', 'early_logins']
mu = train[numeric].mean()
sd = train[numeric].std(ddof=0)
assert sd.gt(0).all()
categories = ['CS', 'Economics', 'Psychology']

def vectors(part):
    z = ((part[numeric] - mu) / sd).to_numpy()
    hot = np.array([
        [int(p == c) for c in categories]
        for p in part['programme']
    ])
    return np.column_stack([z, hot])

train = train.sort_values('id').reset_index(drop=True)
train_vectors = vectors(train)

def predict(part):
    labels = []
    for row_id, v in zip(part['id'], vectors(part)):
        distances = np.linalg.norm(train_vectors - v, axis=1)
```

```
# argmin picks the first minimum; IDs were sorted.
j = int(distances.argmin())
label = train.loc[j, 'outcome']
print(row_id, distances, train.loc[j, 'id'], label)
labels.append(label)
return labels

test_predictions = predict(parts['test'])
new_predictions = predict(parts['new'])
actual = parts['test']['outcome'].to_numpy()
error = np.mean(np.array(test_predictions) != actual)
print('Test error:', error, '| test n:', len(actual))
```

What to explain to the TA: median 90 came only from observed training minutes; scaler parameters came from completed training rows; Physics follows the saved ignore policy; outcomes were excluded from inputs; ratios involving replacements remain marked; a correct prediction on two test students is not proof of broad reliability.

7. Week 5: EDA, confounding and Simpson's paradox

Source: W5 PDF pp. 3-51, 53. The lecture uses both a constructed 200-student support example and a synthetic 4,665-registration course dataset. They are different tables with different units.

7.1 Structured exploration

EDA is a cycle: ask a question, check row meaning and data quality, describe distributions, compare relevant groups, interpret and revisit assumptions. A gallery of unrelated plots does not substitute for that reasoning.

For every figure, identify its population, axes and units, denominator, pattern, possible alternative explanation and limitation. Explain what would change the conclusion. If EDA is being used to choose predictive features, develop them inside training data and protect final test outcomes.

7.2 Counts, rates and denominators

Six passes among ten students is 60%; eight among twenty is 40%. A larger pass count does not imply a higher pass rate. The formula needs eligible units in its denominator.

Suppose A passes three registrations and B fails one. Registration pass rate is $3/4=75\%$. The proportion of students passing at least one course is $1/2=50\%$. Both are valid answers to explicitly different questions. A claim about students requires a student-level outcome definition, such as at least one pass, all courses passed or mean course score.

A mean of group means is not necessarily the overall mean. Use valid group sizes as weights if reconstructing a pooled observed numerical mean. For rates, sum successes and eligible denominators, or equivalently weight subgroup rates by their shares of the relevant denominator.

7.3 Percentage points versus relative percentage

W5's course data contain 3,059 passes among 4,108 full-time registrations (74.464%) and 429 among 557 part-time registrations (77.020%). The difference is about **2.555 percentage points**, or 2.6 points rounded to one decimal.

The relative increase is $(77.020-74.464)/74.464$ times 100, about 3.432%. State which comparison you report. If you subtract only the already rounded 77.0 and 74.5, you get 2.5; retain raw fractions until the final rounding.

7.4 Unknown outcomes produce an honest range

In 100 registrations with 60 passes, 20 failures and 20 unknowns, the observed-outcome rate is $60/80=75\%$. For all 100, the pass rate is at least $60/100=60\%$ and at most $80/100=80\%$. Unknowns do not automatically become failures or successes.

Use the denominator appropriate to the stated question. Label 75% as the rate among known outcomes, and show missingness and bounds when discussing all registrations. Such bounds express logical possibilities from unknown outcomes, not statistical confidence intervals.

7.5 Match plots to variables

Variables and question	Useful view and what to inspect
One category	Bar chart or count/proportion table; name denominator
One numeric variable	Histogram; centre, spread, tails, peaks, gaps and flags
Numeric variable by category	Side-by-side boxplots or distributions with group sizes
Two numeric variables	Scatter plot; direction, form, strength, clusters and unusual points
Two categories	Cross-tabulation or conditional percentages with explicit denominators
A quantity over time	Ordered time plot; use the relevant time unit and watch missing periods

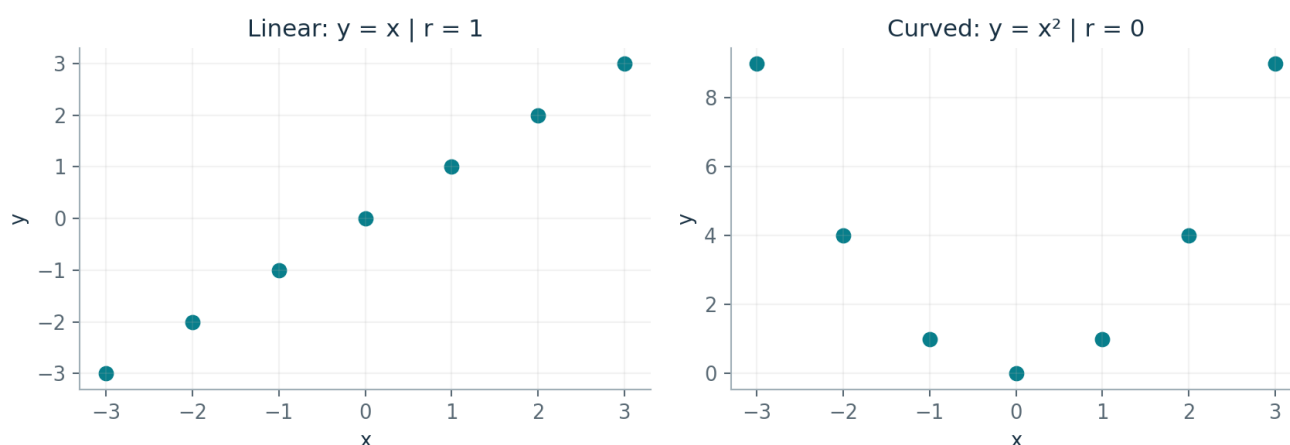
In the histogram mini-example [42,47,53,58,61,64,65,78], bins 40-49,50-59,60-69,70-79 have counts 2,2,3,1. Counts add to 8. In code, define interval boundaries unambiguously so a boundary value is counted once, and ensure the final bin includes the permitted maximum.

A scatter plot can suggest a positive association while individual outcomes still vary widely. A selected subset of 50 plotted points is not the same evidence as a full-cohort correlation computed on 1,200 students. Label the plotted sample and the statistic's sample separately.

7.6 Correlation captures a linear pattern

Pearson correlation r is unitless and ranges from -1 to 1 when both variables vary. Its sign indicates linear direction and magnitude summarises linear association. It does not express the slope in original units, establish causation or capture every kind of relationship.

For symmetric x values -3,-2,-1,0,1,2,3 with $y=x$ squared, the relationship is perfectly deterministic and U-shaped, yet linear correlation is 0. Plot before treating a zero correlation as no relationship. A constant variable makes the usual correlation denominator zero.



Additional reproduction of the Week 5 teaching pattern: linear correlation is 1 for $y=x$ and 0 for the symmetric U-shaped $y=x$ squared. Zero linear correlation is compatible with a strong relationship.

7.7 Simpson's paradox: calculate all three views

The constructed support study is observational. Readiness is measured before support; participation is not randomly assigned.

Readiness	Support: passes/total	No support: passes/total
Lower	48/80 = 60%	10/20 = 50%
Higher	18/20 = 90%	64/80 = 80%

Readiness	Support: passes/total	No support: passes/total
Pooled	66/100 = 66%	74/100 = 74%

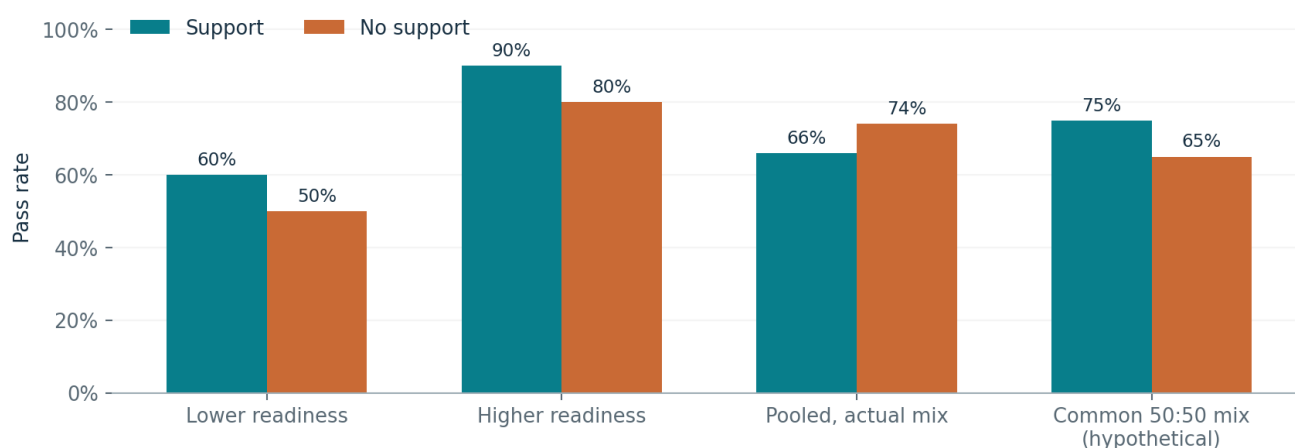
Within each readiness level, support has a 10 percentage-point advantage. Overall, support has an 8 percentage-point disadvantage. There is no arithmetic contradiction: the support group contains 80% lower-readiness students, while the no-support group contains 20%.

The weighted calculations are:

- Support: $0.80 \times 0.60 + 0.20 \times 0.90 = 0.66$.
- No support: $0.20 \times 0.50 + 0.80 \times 0.80 = 0.74$.

Simply averaging 60% and 90% gives 75%, which assumes equal subgroup weights and is not the observed support-group rate. Correct pooling uses subgroup size divided by group total; weights sum to 1.

At a hypothetical common 50:50 readiness mix, support gives $0.5 \times 60\% + 0.5 \times 90\% = 75\%$, and no support gives $0.5 \times 50\% + 0.5 \times 80\% = 65\%$. These standardised rates answer a common-composition question. They are calculated scenarios, not newly observed students and not automatically causal effects.



Rates from W5's constructed support example. The pooled comparison reverses because readiness weights differ; a hypothetical common mix separates that composition effect.

Simpson's paradox: a pooled association can have the opposite direction from associations within groups. To explain it, name the grouping variable, calculate conditional rates, show different weights, and state the claim limit. Do not merely write the label.

7.8 The course-registration example

Course group	Full-time passes/registrations	Part-time passes/registrations
Quantitative service: MA110, CH110, ST110	305/1,027 = 29.7%	7/38 = 18.4%
Other nine courses	2,754/3,081 = 89.4%	422/519 = 81.3%
Pooled	3,059/4,108 = 74.5%	429/557 = 77.0%

Full-time leads in both broad course groups. Part-time registrations put much more weight on the higher-pass group: 519/557 about 93.2%, compared with 3,081/4,108 = 75% for full-time. That composition explains how the pooled ranking can reverse.

Broad groups can still conceal differences. The slides report that part-time has lower observed rates in 11 of 12 individual courses, not all 12. EC102 is an exception. A1/9 small cell is unstable: one extra pass changes 11.1% to 22.2%. Always keep counts beside percentages.

The source's pass definition is final grade at least 50. Different pass thresholds, missing outcomes or eligibility definitions would change the comparison. Re-check these before comparing another dataset.

7.9 Why stratification does not prove causation

Readiness may influence both support participation and final assessment. Grouping by baseline readiness checks one plausible source of confounding. Motivation, time, unmeasured preparation and selection may still differ within groups. Higher conditional rates remain observational evidence unless a suitable design and assumptions support a causal interpretation.

Timing matters for adjustment too. Baseline readiness precedes support; tutoring attendance measured after support might itself be changed by the intervention. Automatically grouping by every available variable can distort the causal question. More detailed subdivision is not automatically more valid, and tiny cells can become unstable.

The descriptive pooled result can be relevant to the actual observed mix, while a conditional result answers a different comparison. Choose the view based on the question and data-generating process, rather than treating either the pooled or stratified result as universally correct.

7.10 A compact EDA implementation pattern

This pattern assumes the course CSVs have been loaded and `passed` is a Boolean/0-1 outcome. The CSV files were not attached, so use it as a code-reading template rather than as a claim about an executed dataset here.

```
joined = enrolments.merge(
    students[['student_id', 'study_mode']],
    on='student_id', how='left', validate='many_to_one'
)
assert len(joined) == len(enrolments)

def rates(group_columns):
    out = joined.groupby(group_columns)['passed'].agg(
        registrations='size', known='count', passes='sum'
    )
    out['missing'] = out['registrations'] - out['known']
    denom = out['known'].where(out['known'].gt(0))
    out['known_pass_rate'] = out['passes'] / denom
    return out

pooled = rates(['study_mode'])
within_course = rates(['course_code', 'study_mode'])
```

An all-missing group has no known-outcome rate; do not call its zero observed successes a 0% pass rate. Inspect matching and missing group labels as well. If evaluating predictions for new students, keep each student's registrations in one split.

Lab 5's lecture handoff asks for an explicit question and population, checked join keys and missing n, pooled versus within-course rates, and an observation, explanation and limitation. No separate Lab 5 notebook was attached.

8. Calculation and code reference

8.1 Formulas you must be able to interpret

Quantity	Formula and required context
Observed mean	Sum of observed values / valid n; state units and population
Weighted mean/rate	Sum of group weight times group value; state what defines weights
Sample variance	$\text{Sum}((x - \text{mean})^2)/(n-1)$; units squared
Population-style variance	$\text{Sum}((x - \text{mean})^2)/n$; use the requested convention
SD	Square root of variance; original units
Range	Maximum-minimum; distinguish width from endpoints
IQR	$Q3 - Q1$; state quartile convention
Review fences	$Q1 - 1.5 \cdot \text{IQR}$ and $Q3 + 1.5 \cdot \text{IQR}$; not deletion rules
Z-score	$(x - \text{reference mean})/\text{reference SD}$; reference SD must be positive
Min-max	$(x - \text{training min})/(\text{training max} - \text{training min})$
Log transform	$\ln(1+x)$, used here for nonnegative counts
Euclidean distance	Square root of sum of squared coordinate differences
Response rate	Replies/invited; use group-specific invited denominators
Missing fraction	Missing/eligible; distinguish cells from absent people
Percentage-point gap	$100 \cdot (p_A - p_B)$ when p values are proportions
Relative percent difference	$100 \cdot (p_A - p_B)/p_B$; requires a stated nonzero reference
Classification error	Wrong labels/evaluated cases
Accuracy	$\text{Correct}/\text{evaluated} = 1 - \text{error}$ under the same counted cases
Sensitivity mean	$[\text{Observed sum} + \text{missing count} \cdot \text{assumed missing mean}]/\text{eligible } N$

8.2 Read the pandas operations in plain English

Code pattern	Meaning and common trap
<code>df.shape, len(df)</code>	Table dimensions or number of rows; not necessarily people
<code>s.nunique(), s.is_unique</code>	Distinct values or uniqueness; inspect missing keys too
<code>s.count()</code> versus <code>s.size</code>	Nonmissing values versus entries
<code>s.isna() / s.notna()</code>	Missingness mask; an empty string may need separate normalisation
<code>s.eq('')</code>	Exact blank strings only
<code>s.value_counts()</code>	Category frequencies; defaults may omit missing categories
<code>s.mode()</code>	Most frequent value(s); can return several
<code>s.mean(), s.median()</code>	Summaries generally skip missing values; report valid n
<code>s.std(ddof=1)</code>	Sample SD; use 0 for the W4 population-style scaler
<code>s.quantile(.25)</code>	A quartile under the software's stated convention
<code>groupby(...).agg(...)</code>	Split into groups and summarise; specify the row unit first
<code>merge(..., validate=...)</code>	Join with a cardinality check; also check unmatched keys
<code>to_numeric(errors='coerce')</code>	Parse numbers; unparsed entries become missing
<code>to_datetime(format=...)</code>	Parse using the source's date convention
<code>fillna(saved_value)</code>	Fill with a specified replacement; does not recover truth
<code>between(a,b)</code>	Check a numeric range; cannot establish source accuracy
<code>assert condition</code>	Stop if the condition fails; only checks what you wrote

8.3 Python ingredients appearing in your labs

`range(12)` supplies integers 0 through 11. In `f'{101+i:05d}'`, `05d` means an integer padded to five digits, so the first ID is '00101'. A list comprehension repeats an expression for each item.

`rng.integers(55,91,12)` draws 12 integers from 55 through 90: the upper bound is excluded.

`loc` selects by label or Boolean condition; `iloc` selects by integer position. `base.iloc[[1]]` returns a one-row table, while a single index can return a row-like Series. `copy(deep=True)` creates an independent copy for the scalar-column records used in the lab. `concat` stacks tables; it does not deduplicate them. `ignore_index=True` rebuilds row index labels, not student IDs.

`np.nan` and `pd.NA` represent missing values in different dtype contexts. Test missingness with `isna`; equality with `NaN` is not a reliable missing-value test. In the original lab code, `np.select` assigns the first matching status rule. Checking original blanks before parser failures preserves their different provenance.

An f-string formats output. A seed makes the random construction reproducible given the same algorithm and sequence of draws. The seed is not a model parameter and does not establish that a case is representative. A case code identifies the generated version without being the student's exam result.

9. Practice exam: 100 marks

This is additional practice designed from the uploaded topics, not an official past paper or a forecast of the real questions. Suggested time: 90 minutes without notes, followed by 30 minutes of correction. Use a calculator if allowed by your actual exam rules.

Q1. Framing and claims (10 marks)

A manager says, "Students who use the platform more get better grades, so we should force everyone to log in."

- Write an analysable question with population, measured variable, time period and purpose. (4)
- Identify the descriptive evidence and the stronger causal claim. (2)
- Give one plausible pre-existing common cause and explain its role. (2)
- State one limitation of login counts as an engagement measure. (2)

Q2. Rows and joins (10 marks)

A registry has 200 unique student IDs. Enrolments has 740 course-registration rows. All enrolment IDs match the registry. One registry ID is accidentally copied; that student has 4 registrations.

- State the row definition in each table and a suitable registration key under one registration per course in the period. (3)
- Predict left-join size before and after the registry copy is introduced. Explain. (3)
- State the appropriate merge cardinality check. (2)
- Explain why a registration pass rate is not automatically a percentage of students passing. (2)

Q3. Summary calculation (10 marks)

Scores are 20,40,60,80 points.

- Compute mean and median. (2)
- Show deviations and sum of squared deviations. (2)
- Calculate sample variance and sample SD with units. (3)
- Calculate the population-style SD and explain the divisor difference. (3)

Q4. Shape, fences and reference groups (10 marks)

A dataset has $Q1=20$, $Q3=24$, and largest observed value inside the upper fence $=28$. Separately, a student scores 87 against a reference mean 75 and SD 6.

- Calculate IQR and both fences. (3)
- State the upper whisker and why it differs from the fence. (2)
- Calculate and interpret z with the stated reference. (2)
- Can you infer a percentile from z alone? Explain. (1)
- Do equal mean and median prove normality? Give a counterexample idea. (2)

Q5. A cleaning decision (10 marks)

Six registry rows represent: A score 70; a confirmed copy of A; B score 60; verified correction B score 80; C age 220 score 90; D verified age 47 with blank score. One final row should represent one student.

- Give actions for the copy, correction, age 220 and age 47. (4)
- Give final row count, valid score count and observed mean. (3)
- Name one accuracy check and one limitation about the missing score. (3)

Q6. Missingness and representation (10 marks)

Day:80 invited,72 replies, reply mean 4. Evening:20 invited,8 replies, reply mean 2.

- Calculate each response rate and Evening's invitation and reply shares. (4)
- Compute the pooled observed mean with valid n. (2)
- State the condition within each class required for the MAR story. (2)
- If all Day students average 4 and the 12 missing Evening ratings average 1, does the all-student mean meet 3.5? (2)

Q7. Transformations (10 marks)

Training logins are 10,20,30. A future record has 40. Programme order is[CS,Economics,Psychology].

- Apply the saved min-max rule to 40 and explain why refitting is inappropriate. (3)
- Give $\log_{1p}(0)$ and $\log_{1p}(9)$, and say what the transform preserves. (2)
- EncodeCS and unseenPhysics under ignore-unknown. (2)
- Explain why raw Euclidean distance can be dominated by minutes while a threshold tree often does not need scaling. (3)

Q8. Pipeline and timing (10 marks)

At the end of W4, you consider a W3 quiz score released in W5 and an exam date announced in W1 for W8. Training scores[60,60,80,80] and logins[2,2,6,6] supply saved rules. A future student has score 78, missing logins, programmeCS.

- Decide availability of the two candidate inputs. (2)
- Give the saved login replacement and score/login scaling parameters. (3)
- Produce the final five-position vector in the lecture order. (2)
- Explain why fitting the scaler on train plus test is leakage and give the correct order. (3)

Q9. EDA and performance limits (10 marks)

There are 60 passes,20 failures and 20 unknown outcomes. A separate scatter plot follows $y=x$ squared with symmetric x values. A classifier gets both of two held-out cases correct.

- Give the known-outcome pass rate and the all-registration lower/upper bounds. (4)
- Explain why correlation 0 does not contradict the curve. (2)
- Give test error and explain why this is weak performance evidence. (2)
- State why exploring final test outcomes to choose features compromises evaluation. (2)

Q10. Simpson's paradox (10 marks)

Lower readiness: support 48/80, no support 10/20. Higher readiness: support 18/20, no support 64/80.

- Compute both within-group comparisons and both pooled rates. (4)
- Show the weights explaining the pooled reversal. (2)
- Compute the rates at a common 50:50 mix. (2)
- Does either comparison prove a causal effect of support? Explain. (2)

10. Practice exam: answers and marking guide

These mark allocations are for self-assessment. They are not the lecturer's rubric. Award reasoning marks only when the answer explains the mechanism, not just when it names a term.

Answer 1

An acceptable framing: "Among first-year students this term, how do recorded W1-W4 logins relate to final grades, and can that evidence identify groups for a support review?" Name the population, variables, time and decision for 4 marks. The observed association is descriptive; requiring logins to raise grades is causal (2). Prior preparation may influence both platform use and grades (2). Login counts can miss offline learning or include unproductive sessions, so they incompletely measure engagement (2). A causal version would require a suitable design rather than the same association restated.

Answer 2

Registry row=student; enrolment row=student-course registration; candidate key=(student_id,course_code) within the defined period (3). With unique registry keys, the left join has 740 rows. The copied student matches each of their 4 registrations twice, adding 4 rows for 744 (3). Use validate='many_to_one'; duplicate right keys should cause the check to fail (2). A student with several courses contributes several registration outcomes, so define and aggregate a student-level success criterion before reporting people (2).

Answer 3

Mean= $200/4=50$; median= $(40+60)/2=50$ (2). Deviations=-30,-10,10,30; squared deviations=900,100,100,900; sum=2,000 (2). Sample variance= $2,000/3=666.667$ points squared, SD about 25.820 points (3). Population-style variance= $2,000/4=500$, SD about 22.361 points. The sample variance convention corrects for estimating the mean; the whole listed-population calculation uses n (3). Do not say SD has squared units.

Answer 4

IQR=4; lower fence= $20-6=14$; upper fence= $24+6=30$ (3). Upper whisker=28, the extreme observed value within the fence, not 30 (2). $z=(87-75)/6=2$: two SD above this reference mean (2). A percentile needs the empirical distribution or a stated model (1). A symmetric two-peak distribution can have equal mean and median without being normal (2).

Answer 5

Remove the confirmed extra A row, retain the verified B correction, flag C's invalid age while retaining usable fields, and retain verified D (4). Final rows 4; observed scores 70,80,90; valid n = 3; mean 80 points (3). Compare scores with the exact marked exam/source; the blank D score prevents claiming that all four average 80 and may make observed cases unrepresentative (3). A range assertion alone does not establish accuracy.

Answer 6

Day response $72/80=90\%$; Evening $8/20=40\%$; Evening invited $20/100=20\%$; Evening replies $8/80=10\%$ (4). Pooled observed mean= $(288+16)/80=3.8$, valid n = 80 (2). Within each known class, response probability must be independent of satisfaction for the stated MAR story (2). The alternative all-student mean= $(320+16+12)/100=3.48$, below 3.5 (2). Different observed response rates do not prove the MAR assumption.

Answer 7

Saved min/max 10 and 30 give $(40-10)/(30-10)=1.5$. Apply the existing reference; refitting lets future data alter it (3). $\log_{10}(0)=0$; $\log_{10}(9)=\ln 10$ about 2.303; it preserves order (2). $CS=[1,0,0]$; $Physics=[0,0,0]$ under the chosen policy (2). Large numeric minute gaps dominate squared raw distances; a threshold tree uses numerical ordering, whose equivalent thresholds can be retained under monotone scaling (3).

Answer 8

Quiz score unavailable at W4; announced future exam schedule available (2). Saved login median 4; score mean 70/SD 10; login mean 4/SD 2 using ddof0 (3). $Vector[0.8,0,1,0,0]$ (2). Held-out rows must not influence learned preprocessing. Split first, fit preprocessing on training, transform both with saved rules, fit model on training and evaluate fixed predictions on test (3).

Answer 9

Known pass rate $60/80=75\%$; full-population bounds $60/100=60\%$ and $80/100=80\%$ (4). Pearson r measures linear association; a symmetric U-shape can give $r=0$ despite a deterministic relationship (2). Error $0/2=0\%$, but two test observations give weak evidence of reliable generalisation (2). Feature selection using final test outcomes lets that set influence the procedure, so it is no longer an independent final assessment (2).

Answer 10

Lower: 60% versus 50% ; higher: 90% versus 80% ; pooled support 66% , no support 74% (4). Support weights $0.8/0.2$ give 66% ; no-support $0.2/0.8$ give 74% (2). Common 50:50 rates 75% and 65% (2). Neither establishes causation: nonrandom participation and unmeasured within-group differences may remain; the common mix removes a specified composition difference, not every source of bias (2).

11. Revision traps, writing templates and readiness

11.1 Twenty distinctions that protect marks

1. Rows are not automatically people; the key follows the row definition.
2. A repeated ID can be a copy, a correction or a legitimate repeated observation.
3. Correct arithmetic can still use the wrong denominator.
4. Size includes missing entries; count excludes them.
5. A valid numeric range does not establish source accuracy.
6. A missing value is not zero and not automatically not-applicable.
7. MCAR/MAR/MNAR describe mechanisms; response counts alone do not identify them.
8. Equal response rates do not guarantee an unbiased comparison.
9. Recovering a verified observation differs from imputing an estimate.
10. Filling cells does not increase the number of observed responses.
11. An IQR flag is not a deletion instruction; a whisker is not a fence.
12. Mean/median agreement and near-zero skewness do not prove normality.
13. Sample and population-style SD use different divisors by design.
14. A z-score needs a reference group; it is not automatically a percentile.
15. A future min-max value may lie outside 0-1.
16. A past event's result can be unavailable; a future schedule can already be known.
17. Fit learns from training; transform applies the saved rule.
18. Perfect training accuracy and tiny-test accuracy do not prove future reliability.
19. Pooled rates are weighted averages; changed composition can reverse a comparison.
20. Grouping and standardising do not automatically establish a causal effect.

11.2 Reporting templates

Descriptive summary: "Among [valid n] observed [units] from [population/period], [quantity] had [centre] and [spread] in [measurement units]. The distribution was [shape]. [Missing count or other specific limitation] limits [the intended conclusion]."

Cleaning decision: "I [action] [record/value] because [dictionary/source evidence]. This changes [rows/cells/values] by [count]. I retained [raw/provenance] and [uncertainty]. The alternative would [specific consequence]."

Transformation: "I fitted [parameters] using [training rows]. For [new value], [substitution] gives [output]. The same rule is reused for test/future records; [availability or interpretation limit]."

Comparison: "[Group A] had [successes/denominator and rate], compared with [Group B]. The difference is [percentage points or other unit]. [Composition/selection factor] may explain part of it, so this does not establish [causal claim]."

Simpson explanation: "Within [strata], [direction] holds, whereas the pooled direction is [opposite]. [Group A] places more weight on [lower/higher-rate stratum], with weights [values]. These answer different composition-specific questions; [remaining causal limit]."

Use these as structures, not scripts to paste unchanged. Marks depend on the evidence and quantities in the question you are given.

11.3 Precision notes and slide inconsistencies

The source material contains useful reminders that should not be overgeneralised. Use the precise statement in calculations and explanations.

- **W2 PDF p. 15:** the heading says an impossible value moves a mean by 34 years, while the figure labels show 66.0 and 95.9, a difference of 29.9. Do not memorise this inconsistent arithmetic; recompute from the observations supplied in an exam. The valid lesson is that extreme values can move the mean strongly.
- **W2 PDF p. 17:** the caption says eight scores, but the displayed mini-list contains nine. Count the actual observations before locating positions.
- **W2 PDF p. 29:** the $n-1$ correction concerns the expectation of sample variance under sampling assumptions, not every realised sample and not exact unbiasedness of SD.
- **W2 PDF pp. 25,28:** range cannot shrink as observations are added to a fixed list, but may stay constant. SD and IQR can change; the distinction is their different behaviour, not that they never move.
- **W2 PDF pp. 45,49:** interpreting $z=1$ as about the 84th percentile needs a normal model; the later slide correctly separates distance from rank.
- **W3 PDF pp. 13,50-51:** percentage-to-proportion conversion and 0-100 score conversion target different units. Read the dictionary.
- **W5 PDF pp. 11,53:** the reported 2.6 percentage-point gap uses unrounded fractions; subtracting the displayed one-decimal rates loses precision.
- **Deck numbering:** some W2 printed slide numbers repeat or differ from PDF page order; references in this guide use PDF page positions. The Lab 2 note mentioning slides 2-65 exceeds the 60-page uploaded deck; rely on the actual attached material and ask about any missing revision if needed.

11.4 Readiness check

Before calling a topic learned, solve a new example without notes, explain each denominator and unit, identify a tempting wrong answer, and say what evidence would change your conclusion. Then repeat after a delay. A correct calculation copied from a worked example is not the same as independent mastery.

You are ready for this pack's topics when you can: predict a join's row count; explain a misleading average; calculate SD and quartiles with conventions; distinguish flags from deletions; explain missingness within groups; reconstruct a sensitivity bound; transform a future vector from saved parameters; identify availability and test leakage; compute held-out error; and explain a Simpson reversal using weights.

For the actual paper, scan all questions, allocate time in proportion to marks, show substitutions, label denominators and units, and use the last portion of your time to check omissions and overclaims. If asked for a choice, commit to the supported choice and state the condition that makes it defensible.

11.5 How the rest of your earlier syllabus connects

Your earlier syllabus continues with visual communication (Week 6), probability and Bayes (7), inference and A/B design (8), regression (9), classification and fairness (10), clustering and PCA (11), and time series, text, dashboards and responsibility (12). The current guide supplies their foundations: meaningful measurements, data quality, honest evaluation and cautious interpretation. Detailed preparation for those later topics needs their actual lecture and lab materials.

12. Source map and coverage checklist

All page references below are one-based PDF positions. Notebook sections use the visible headings rather than internal cell numbers. Course examples are synthetic. Added practice questions and visualisations are clearly labelled; they are not official exam questions.

Label	Uploaded source	Main material used
W1	week-1-lecture.pdf, 60 pages	Framing, claims, rows, measurement, sampling, proxies, joins, ethics and assessment
L1	Lab_01_Rows_Joins_Missingness_Student(1).ipynb, 23 cells	Two-table merge, pass-rate grouping, denominator and reasoning rubric
W2	Week_2_Describing_Data_SLIDES.pdf, 60 pages	Centre, spread, shape, quartiles, normal approximation and z-scores
L2	Lab_02_Describing_Data_Student.ipynb, 34 cells	Personalised 240-student case, hand calculations, conventions, plots and transfer tasks
W3	data-science-fundamentals-week-3.pdf, 56 pages	Quality, duplicate evidence, missingness, sensitivity, outliers and audit
W4	data-science-fundamentals-week-4(1).pdf, 43 pages	Transformations, encoding, timing, pipelines, leakage and honest error
L4	data-science-fundamentals-week-4-lab-instructions.pdf, 7 pages	Seven fixed practical tasks and the combined nearest-student exercise
W5	data-science-fundamentals-week-5(1).pdf, 53 pages	EDA, plots, conditional rates, weighting, confounding and Simpson's paradox

Topic-to-source locator

Topic	Where to return
Applied exam format and weights	W1 pp. 19-20
Question framing and claim ladder	W1 pp. 22-29
Lifecycle, measurement and rows	W1 pp. 30-39
Frames, proxies, joins and ethics	W1 pp. 42-53; L1 Parts 1-5
Summary choice and hand calculations	W2 pp. 10-30; L2 Sections 2-3
Distribution, bins and normal approximation	W2 pp. 33-43; L2 Sections 4-6
Z-score reference, rank and specification	W2 pp. 45-50; L2 Section 7
Typical use versus total workload	W2 p. 58; L2 Section 8
Quality and duplicate reasoning	W3 pp. 6-21
Missingness and representation	W3 pp. 22-39,45
Outliers, logs and checks	W3 pp. 40-54
Distance, scaling and log	W4 pp. 4-16; L4 Tasks 1-2
Encoding and engineered features	W4 pp. 17-22; L4 Task 3
Availability and leakage	W4 pp. 24-28,33-35; L4 Task 5
Pipeline vectors and evaluation	W4 pp. 29-40; L4 Tasks 4,6-7
EDA, rates, plots and unknown outcomes	W5 pp. 5-24
Support example and common-mix rates	W5 pp. 26-37
Course mix, causation and test boundary	W5 pp. 38-51,53

Final study principle: calculations earn meaning from their population, units, timing, denominator and evidence. Practise supplying those details until they become automatic.